

Non-reversible Metropolis-Hastings

Joris Bierkens¹ 

Received: 12 March 2014 / Accepted: 23 July 2015 / Published online: 18 August 2015
© The Author(s) 2015. This article is published with open access at Springerlink.com

Abstract The classical Metropolis-Hastings (MH) algorithm can be extended to generate non-reversible Markov chains. This is achieved by means of a modification of the acceptance probability, using the notion of vorticity matrix. The resulting Markov chain is non-reversible. Results from the literature on asymptotic variance, large deviations theory and mixing time are mentioned, and in the case of a large deviations result, adapted, to explain how non-reversible Markov chains have favorable properties in these respects. We provide an application of NRMH in a continuous setting by developing the necessary theory and applying, as first examples, the theory to Gaussian distributions in three and nine dimensions. The empirical autocorrelation and estimated asymptotic variance for NRMH applied to these examples show significant improvement compared to MH with identical stepsize.

Keywords Markov Chain Monte Carlo · MCMC · Metropolis-Hastings · Non-reversible Markov processes · Asymptotic variance · Large deviations · Langevin sampling

Mathematics Subject Classification 65C40 · 60J20

1 Introduction

The Metropolis-Hastings (MH) algorithm [Metropolis et al. \(1953\)](#), [Hastings \(1970\)](#) is a Markov chain Monte-Carlo (MCMC) method of profound importance to many fields of mathematics such as Bayesian inference and statisti-

cal mechanics [Diaconis and Saloff-Coste \(1998\)](#), [Diaconis \(2008\)](#), [Levin et al. \(2009\)](#). The applicability of MH to a particular computational problem depends on the efficiency of the Markov chain that is generated by the algorithm. The chains generated by the classical Metropolis-Hastings algorithm are *reversible*, or, in other words, satisfy *detailed balance*; in fact, this reversibility is instrumental in showing that the resulting chains have the right invariant probability distribution.

However, non-reversible Markov chains may have better properties in terms of mixing behavior or asymptotic variance. This can be shown experimentally in special cases [Suwa and Todo \(2010\)](#), [Turitsyn et al. \(2011\)](#), [Vucelja \(2014\)](#), theoretically in special cases [Diaconis et al. \(2000\)](#), [Neal \(2004\)](#), and in fact, also in general [Sun et al. \(2010\)](#), [Chen and Hwang \(2013\)](#), with respect to asymptotic variance. See also [Rey-Bellet and Spiliopoulos \(2014\)](#) for improved asymptotic variance of non-reversible diffusion processes on compact manifolds.

There exist two basic approaches to the construction of non-reversible chains from reversible chains: one can ‘lift’ the Markov chain to a larger state space [Diaconis et al. \(2000\)](#), [Neal \(2004\)](#), [Turitsyn et al. \(2011\)](#), [Vucelja \(2014\)](#), or one can introduce non-reversibility without altering the state space [Sun et al. \(2010\)](#). In continuous spaces, the *hybrid* or *Hamiltonian Monte Carlo* [Horowitz \(1991\)](#), [Neal \(2011\)](#) is closely related to the lifting approach in discrete spaces. Other noteworthy publications on non-reversible Markov chains are [Wilmer \(1999\)](#), [Geyer and Mira \(2000\)](#).

In this paper we consider the second type of creating non-reversibility, i.e. without augmenting the state space. In discrete spaces this can, in principle, be achieved by changing transition probabilities (see [Remark 2.2](#)). However this may be have computational disadvantages, since it requires access to all transition probabilities. Furthermore, there is no such

✉ Joris Bierkens
j.bierkens@warwick.ac.uk

¹ Department of Statistics, University of Warwick,
Coventry CV4 7AL, UK

analogue in continuous spaces (crudely speaking, because all transition probabilities to specific states are zero).

To remedy these issues, in this paper MH is extended to ‘non-reversible Metropolis-Hastings’ (NRMH) which allows for non-reversible transitions. The main idea of this paper is to modify the acceptance ratio, which is further discussed in Sect. 2. For pedagogical purposes, the theory is first developed for discrete state spaces. It is shown how the acceptance probability of MH, can be adjusted so that the resulting chain in NRMH has a specified ‘vorticity’, and therefore, will be non-reversible. Any Markov chain satisfying a symmetric structure condition can be constructed by NRMH, which establishes the generality of the algorithm. Theoretical advantages of finite state space non-reversible chains in terms of improved asymptotic variance and large deviations estimates are briefly mentioned in Sect. 3. In particular we recall a result from Sun et al. (2010) that adding non-reversibility decreases asymptotic variance. Also we present a variation on a result by Rey-Bellet and Spiliopoulos (2014) on large deviations from the invariant distribution.

As mentioned above, for continuous state spaces it was so far not clear how general non-reversible Markov chains (i.e. discrete time, for arbitrary target density) could be constructed. One of the main advantages of NRMH is that it also applies in the setting of continuous state spaces, and thus provides a partial solution to this problem, as will be discussed and verified experimentally in Sect. 4. In particular we implement a non-reversible version of the Metropolis Adjusted Langevin Algorithm (MALA) for Gaussian multivariate target distributions. Finally, conclusions and directions of further research are discussed in Sect. 5.

1.1 Notation

We will consider both finite and infinite-dimensional vectors and matrices. The constant vector with all elements equal to 1 will be denoted by $\mathbb{1}$; the dimensionality of $\mathbb{1}$ should always be clear from the context. Similarly the identity matrix of any dimension will be denoted by I . For sets V, S , with $V \subset S$ the indicator function of V is denoted by $\mathbb{1}_V : S \rightarrow \mathbb{R}$. The transpose of a matrix A is denoted by A' . The Euclidean vector norm on \mathbb{R}^n , as well as its induced matrix norm, will be denoted by $\|\cdot\|$. For a matrix $A \in \mathbb{R}^n$, the spectrum is denoted by $\sigma(A)$. The spectral bound and spectral radius of A are denoted by $s(A) = \max\{\operatorname{Re} \lambda : \lambda \in \sigma(A)\}$ and $\tau(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$, respectively.

2 Metropolis-Hastings generalized to obtain non-reversible chains

As a preliminary to non-reversible Metropolis-Hastings, we require the notion of *vorticity matrix*, which is introduced

in Sect. 2.1. The classical Metropolis-Hastings algorithm, discussed in Sect. 2.2, is extended using the notion of vorticity matrix to a non-reversible version in Sect. 2.3.

2.1 Non-reversible Markov chains and vorticity

Let $P = (P(x, y))$ denote a matrix of transition probabilities of a Markov chain on a finite or countable state space S . A *distribution* on S is a vector with positive elements in $L^1(S)$, and is not necessarily normalized, i.e. it is not necessarily the case that $\sum_{x \in S} \pi(x) = 1$. If π is a distribution such that $\sum_{x \in S} \pi(x) = 1$, then we call π a *probability distribution*. We will always assume that $\pi(x) > 0$ for all $x \in S$. A (probability) distribution π on S is said to be an *invariant (probability) distribution* of P if $\pi'P = \pi'$, i.e. $\sum_{x \in S} \pi(x)P(x, y) = \pi(y)$ for all $y \in S$. A distribution π on S is said to satisfy the *detailed balance condition* with respect to P if $\operatorname{diag}(\pi)P = P' \operatorname{diag}(\pi)$, i.e. $\pi(x)P(x, y) = P(y, x)\pi(y)$ for all $x, y \in S$. If there exists a distribution π on S that satisfies detailed balance with respect to P , then P is said to be *reversible*. As is well known, and straightforward to check, if π satisfies detailed balance with respect to P , then π is invariant for P . A chain which does not satisfy the detailed balance condition with respect to its invariant distribution is called *non-reversible*. In a certain sense, this is a misnomer: we may obtain a time reversed Markov chain \hat{P} by defining $\hat{P}(x, y) := \frac{\pi(y)P(y, x)}{\pi(x)}$. In fact, \hat{P} is the adjoint of P with respect to the inner product $(\cdot, \cdot)_\pi$, defined by $(f, g)_\pi = \sum_{x \in S} f(x)g(x)\pi(x)$. For further background material on Markov chains the reader is referred to Levin et al. (2009).

Consider a non-reversible Markov chain P . Let $K = \frac{1}{2}(P + \hat{P})$. Then K is a reversible Markov chain with invariant distribution π . We will also define a vorticity matrix Γ as (essentially) the skew-symmetric part of P :

$$\Gamma(x, y) := \pi(x)P(x, y) - \pi(y)P(y, x), \quad x, y \in S, \quad (1)$$

or in matrix notation

$$\Gamma = \operatorname{diag}(\pi)P - P' \operatorname{diag}(\pi).$$

One can think of Γ as (a transformation of) the skew-symmetric part of P : $\Gamma = \operatorname{diag}(\pi)(P - \hat{P})$. The following simple observations are fundamental to this paper.

Lemma 2.1 *Let P be the transition matrix of a Markov chain on S and let π be a distribution on S . Let Γ be defined by (1). Then:*

- (i) Γ is skew-symmetric, i.e. $\Gamma = -\Gamma'$;
- (ii) π satisfies detailed balance with respect to P if and only if $\Gamma = 0$;

(iii) π is invariant for P if and only if $\Gamma \mathbb{1} = 0$.

Proof (i) and (ii) are immediate. As for (iii), note that

$$\begin{aligned} (\Gamma \mathbb{1})(x) &= \sum_{y \in S} [\pi(x)P(x, y) - \pi(y)P(y, x)] \\ &= \pi(x) - \sum_{y \in S} \pi(y)P(y, x), \end{aligned}$$

which is zero for all x if and only if π is invariant for P . \square

In light of Lemma 2.1, a matrix $\Gamma \in \mathbb{R}^{n \times n}$ which is skew-symmetric and satisfies $\Gamma \mathbb{1} = 0$ is called a *vorticity matrix*. If Γ is related by (1) to a Markov chain P with invariant distribution π , it is called the *vorticity of P and π* . It will be a key ingredient in the construction of a non-reversible version of Metropolis-Hastings.

Remark 2.2 A direct way of constructing a non-reversible chain P from a reversible chain K and a vorticity matrix Γ , is by letting $P(x, y) = K(x, y) + \frac{1}{2\pi(x)}\Gamma(x, y)$, provided that P is a probability matrix (i.e. has nonnegative entries). This is discussed in e.g. Sun et al. (2010). In order to make a transition from a state x , one has to compute entries of $K(x, \cdot)$ and $\Gamma(x, \cdot)$ to determine the transition probabilities. This approach has no alternative in uncountable state spaces. This is the main reason for wishing to develop a method that does not depend on the construction mentioned in this remark.

2.2 Metropolis-Hastings

In the Metropolis-Hastings (MH) algorithm a reversible Markov chain P_0 with a given invariant distribution π is constructed. We will assume, mainly for simplicity, throughout this paper that $\pi(x) > 0$ for all $x \in S$. As an ingredient for the construction of P_0 , a Markov chain Q is used, satisfying the *symmetric structure* condition

$$Q(y, x) = 0 \quad \text{whenever} \quad Q(x, y) = 0, \quad x, y \in S. \quad (2)$$

In other words, whenever a transition from x to y has positive probability, the reverse probability also has positive probability. The *Hastings ratio* $R_0(x, y)$ is defined as

$$R_0(x, y) := \begin{cases} \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}, & \text{for all } x, y \in S \text{ for which} \\ & \pi(x)Q(x, y) \neq 0, \\ 1 & \text{otherwise.} \end{cases} \quad (3)$$

With this definition of R_0 , acceptance probabilities are defined as

$$A_0(x, y) := \min(1, R_0(x, y)), \quad (4)$$

and transition probabilities $P_0(x, y)$ are defined by

$$P_0(x, y) := \begin{cases} Q(x, y)A_0(x, y), & x \neq y, \\ Q(x, x) + \sum_{z \neq x} Q(x, z) \times (1 - A_0(x, z)), & x = y. \end{cases} \quad (5)$$

It is a straightforward exercise to show that the chain P_0 has π as its invariant distribution. An important step is the observation that $R_0(x, y) \leq 1$ if and only if $R_0(y, x) \geq 1$, which will be a recurring phenomenon in the sequel.

2.3 Non-reversible Metropolis-Hastings

We will now discuss how this framework can be extended to construct Markov chains that are, in general, non-reversible. Let $\Gamma \in \mathbb{R}^{n \times n}$ be a vorticity matrix, and let Q be the transition matrix of a Markov chain, satisfying (2). Again, $\pi : S \rightarrow (0, \infty)$ is some distribution that is not necessarily normalized and has only positive entries.

We define for $x, y \in S$ the *non-reversible Hastings ratio* as

$$R_\Gamma(x, y) := \begin{cases} \frac{\Gamma(x, y) + \pi(y)Q(y, x)}{\pi(x)Q(x, y)}, & \text{if } \pi(x)Q(x, y) \neq 0, \\ 1 & \text{otherwise,} \end{cases} \quad (6)$$

and let, analogously to MH, the acceptance probabilities A_Γ be

$$A_\Gamma(x, y) := \min(1, R_\Gamma(x, y)). \quad (7)$$

Entries of Γ can be negative. In order to avoid the situation that A_Γ becomes negative, we will explicitly constrain vorticity matrix Γ to satisfy

$$\Gamma(x, y) \geq -\pi(y)Q(y, x) \quad \text{for all } x, y \in S. \quad (8)$$

Note that (8) implies, by skew-symmetry of Γ , that

$$-\pi(y)Q(y, x) \leq \Gamma(x, y) \leq \pi(x)Q(x, y) \quad \text{for all } x, y \in S.$$

In particular, by the symmetric structure condition (2), Γ should have zeroes wherever Q has zeroes. As with Metropolis-Hastings, the transition probabilities $P_\Gamma(x, y)$ are defined by

$$P_\Gamma(x, y) := \begin{cases} Q(x, y)A_\Gamma(x, y), & x \neq y, \\ Q(x, x) + \sum_{z \neq x} Q(x, z) \times (1 - A_\Gamma(x, z)), & x = y. \end{cases} \quad (9)$$

Note that indeed P_Γ is a matrix of transition probabilities. For $\Gamma = 0$, A_Γ and therefore P_Γ reduce to A_0 and P_0 , so that the chosen notation is consistent.

In order to check that the proposed Markov chain has π as its invariant density, we need to verify that Γ , π and P_Γ

are related through (1). As a crucial step, we employ the following lemma, in analogy with Metropolis-Hastings.

Lemma 2.3 *Let Γ be a vorticity matrix, Q a matrix of transition probabilities satisfying (2), π a distribution that is nowhere zero, such that (8) holds. Let R_Γ be as above. Then $R_\Gamma(y, x) > 1$ if and only if $R_\Gamma(x, y) < 1$ for any $x, y \in S$ for which $Q(x, y) \neq 0$.*

Proof Suppose $R_\Gamma(x, y) < 1$, i.e. $\Gamma(x, y) + \pi(y)Q(y, x) < \pi(x)Q(x, y)$. Then

$$\begin{aligned}\pi(x)Q(x, y) + \Gamma(y, x) &= \pi(x)Q(x, y) - \Gamma(x, y) \\ &> \pi(y)Q(y, x),\end{aligned}$$

so that $R_\Gamma(y, x) > 1$, using that Γ is skew-symmetric. The converse direction is analogous. \square

Using the previous lemma, it is now straightforward to show that Γ is the vorticity matrix of (P_Γ, π) .

Lemma 2.4 *Let Q be a Markov chain, Γ a vorticity matrix, and π a distribution on S , such that (2) and (8) are satisfied. Let P_Γ be defined through (6), (7) and (9). Then (1) holds for (P_Γ, π) , i.e.*

$$\Gamma(x, y) = \pi(x)P_\Gamma(x, y) - \pi(y)P_\Gamma(y, x), \quad \text{for all } x, y \in S.$$

Proof If x, y are such that $R_\Gamma(x, y) < 1$. Then $A_\Gamma(x, y) = R_\Gamma(x, y)$, and $A_\Gamma(y, x) = 1$. Therefore

$$\begin{aligned}\pi(x)P_\Gamma(x, y) &= \pi(x)Q(x, y)A_\Gamma(x, y) \\ &= \pi(x)Q(x, y) \left(\frac{\Gamma(x, y) + \pi(y)Q(y, x)}{\pi(x)Q(x, y)} \right) \\ &= \Gamma(x, y) + \pi(y)Q(y, x) = \Gamma(x, y) \\ &\quad + \pi(y)A_\Gamma(y, x)Q(y, x) = \Gamma(x, y) \\ &\quad + \pi(y)P_\Gamma(y, x),\end{aligned}$$

using Lemma 2.3. so that (1) is satisfied for such x and y . The cases in which $R_\Gamma(x, y) = 1$ or > 1 are analogous. \square

Finally, since by the assumption that Γ is a vorticity matrix, $\Gamma \mathbb{1} = 0$, and hence Lemma 2.1 (iii) gives that π is invariant for P_Γ . We have obtained our main result.

Theorem 2.5 *Let Q be a Markov chain, Γ a vorticity matrix, and π a distribution on S that is everywhere positive, such that (2) and (8) are satisfied. Let P_Γ be defined through (6), (7) and (9). Then P_Γ has π as invariant distribution and Γ as its vorticity matrix.*

Remark 2.6 We will refer to a combination (Q, Γ, π) which satisfies (2) and (8) as a *compatible combination*. Especially verifying condition (8) requires some knowledge about π ,

but these do not need to be exact: it will suffice to have access to a lower bound for π . In the proof of Theorem 4.2 the analogue of this condition for continuous spaces is checked as an example.

Remark 2.7 Once we have access to a compatible combination of proposal chain Q , vorticity matrix Γ and target distribution π , the NRMH algorithm has similar favorable properties as MH, in that only local information is required: Q , π and Γ only need to be evaluated at the current and proposed state, and no normalization of π is required. In Sect. 4 this will become even more important when NRMH is applied to a problem in continuous state space.

2.4 General observations on NRMH

The following trivial observation serves to indicate the generality of this approach. It asserts that every Markov chain may be build, in a trivial way, by the described procedure.

Proposition 2.8 *Let P be a Markov chain with invariant distribution π and corresponding vorticity Γ , satisfying (2). If we use $Q = P$ as proposal distribution and Γ as vorticity matrix in the NRMH algorithm with target distribution π , then the resulting Markov chain P_Γ is identical to P .*

Proof It suffices to note, that by $Q = P$ and (1), $A_\Gamma(x, y) = 1$ for all $x, y \in S, x \neq y$. \square

Remark 2.9 If, for some pair $(x, y) \in S \times S$, (8) holds with equality, the transition probability $P_\Gamma(x, y) = 0$ even when $Q(x, y) \neq 0$. Therefore irreducibility of Q does not imply irreducibility of P_Γ , unless we impose the stronger condition:

$$\Gamma(x, y) > -\pi(y)Q(y, x) \quad \text{for all } x, y \in S. \quad (10)$$

As noted in Remark 2.2, one may alternatively construct any non-reversible chain P by ‘adding’ a vorticity matrix to a reversible chain K . We may translate one approach into the other, as follows:

Proposition 2.10 *Consider irreducible transition kernels $Q(x, y)$ and $H(x, y)$, related by $Q(x, y) = H(x, y) + \frac{1}{2\pi(x)}\Gamma(x, y)$, both satisfying the symmetric structure condition (2). Suppose Q, π and Γ satisfy (8). Suppose P_1 is the Markov kernel obtained from NRMH, using Q as proposal chain and Γ as vorticity matrix. Suppose P_2 is the Markov kernel given by $P_2(x, y) = K(x, y) + \frac{1}{2\pi(x)}\Gamma(x, y)$, where K is the classical Metropolis-Hastings kernel obtained by using H as proposal chain (and π as target distribution). Then $P_1 = P_2$.*

Proof We compute, for $x \neq y$, with $Q(x, y) \neq 0$,

$$\begin{aligned}
& \pi(x)P_1(x, y) \\
&= \pi(x) \min \left(1, \frac{\Gamma(x, y) + \pi(y)Q(y, x)}{\pi(x)Q(x, y)} \right) Q(x, y) \\
&= \min (\pi(x)Q(x, y), \Gamma(x, y) + \pi(y)Q(y, x)) \\
&= \min \left(\pi(x)H(x, y) + \frac{1}{2}\Gamma(x, y), \Gamma(x, y) \right. \\
&\quad \left. + \pi(y)H(y, x) + \frac{1}{2}\Gamma(y, x) \right) \\
&= \min (\pi(x)H(x, y), \pi(y)H(y, x)) + \frac{1}{2}\Gamma(x, y) \\
&= \pi(x) \min \left(1, \frac{\pi(y)H(y, x)}{\pi(x)H(x, y)} \right) H(x, y) + \frac{1}{2}\Gamma(x, y) \\
&= \pi(x)P_2(x, y).
\end{aligned}$$

Both P_1 and P_2 have zeros on the off-diagonal entries for which Q has zeros. Since both P_1 and P_2 represent transition probabilities, this also fixes their diagonal elements, which concludes the proof. \square

3 Advantages of non-reversible Markov chains in finite state spaces

Non-reversible Markov chains may offer important computational advantages compared to reversible chains. We will briefly discuss some of these advantages as they apply to finite state space Markov chains. It is to our knowledge an open question how to extend the results of Sects. 3.1 and 3.2 in a generic way to uncountable and continuous state spaces.

3.1 Asymptotic variance

Consider a Markov chain P on S with invariant probability distribution μ . Let $f : S \rightarrow \mathbb{R}$. We say that f satisfies a Central Limit Theorem (CLT) if there is a $\sigma_f^2 < \infty$ such that the normalized sum $n^{-1/2} \sum_{i=1}^n [f(X_i) - \mu(f)]$ converges weakly to a $N(0, \sigma_f^2)$ distribution. In this case σ_f^2 is called the asymptotic variance.

We will in this section work under the assumption that S is finite. In this case, for any $f : S \rightarrow \mathbb{R}$ and irreducible P , a CLT is satisfied [see e.g. Roberts and Rosenthal (2004)]. The following result is obtained in Sun et al. (2010), for a more extensive argument see Chen and Hwang (2013).

Proposition 3.1 *Let K be a transition matrix of an irreducible reversible Markov chain with invariant probability distribution μ . Let Γ be a non-zero vorticity matrix and let $P = K + \frac{1}{2} \text{diag}(\mu)^{-1} \Gamma$ be the transition matrix of an irreducible Markov chain. For any $f : S \rightarrow \mathbb{R}$ and denote by $\sigma_{f,K}^2$ and $\sigma_{f,P}^2$ the asymptotic variances of f with respect to the transition matrices K and P , respectively.*

Then for all $f : S \rightarrow \mathbb{R}$, we have $\sigma_{f,P}^2 \leq \sigma_{f,K}^2$, and there exists an f such that $\sigma_{f,P}^2 < \sigma_{f,K}^2$.

In words, adding non-reversibility decreases asymptotic variance.

3.2 Large deviations

In Rey-Bellet and Spiliopoulos (2014) it is noted that non-reversible diffusions on compact manifolds have favorable properties in terms of large deviations of the occupation measure from the invariant distribution. Inspired by their result, we present a simple (but to our knowledge novel) result in the same direction for finite state spaces.

As in the previous section, assume S is finite. We may transform a discrete time Markov chain on S into a continuous time chain by making subsequent transitions after random waiting times that have independent $\text{Exp}(\lambda)$ distributions. The discrete chain with transition matrix P then transforms into a continuous time chain with generator

$$G(x, y) = \begin{cases} \lambda P(x, y) & \text{if } x \neq y, \\ -\lambda \sum_{z \neq x} P(x, z) & \text{if } x = y. \end{cases}$$

The occupation measure of the resulting Markov process is defined as $L_t = \frac{1}{t} \int_0^t \delta_{X_s} ds$. If G is irreducible, the occupation measure satisfies for large time a large deviation principle with rate function

$$I_G(\mu) = \sup_{u > 0} \left(- \sum_{x \in S} \mu_i \frac{(Gu)(x)}{u(x)} \right), \quad \mu \in \mathcal{P}(S),$$

i.e. informally, for large t , for $A \subset \mathcal{P}(S)$,

$$\mathbb{P}(L_t \in A) \approx \exp \left(-t \inf_{\mu \in A} I_G(\mu) \right).$$

Here $\mathcal{P}(S)$ is the set of probability distributions on S . The rate function satisfies the properties that (i) $I_G \geq 0$, (ii) I_G is strictly convex, and (iii) $I_G(\mu) = 0$ if and only if $\mu = \pi$, where π is the invariant distribution of G . Hence I_G quantifies the probability of a large deviation of the occupation measure from the invariant distribution for large t . See Holander (2000) for further details.

The following result shows that deviations for non-reversible continuous time chains from the invariant distribution are asymptotically less likely than for the corresponding reversible chain.

Proposition 3.2 *Suppose G admits a decomposition of the form $G(x, y) = K(x, y) + \frac{1}{2\pi(x)} \Gamma(x, y)$, where $\pi(x)K(x, y) = \pi(y)K(y, x)$ and $\Gamma(x, y) = -\Gamma(y, x)$ for all $x \neq y$. Then $I_G(\mu) \geq I_K(\mu)$ for all $\mu \in \mathcal{P}(X)$, and the inequality is strict if $\Gamma u^* \neq 0$.*

Proof Assume $\mu(x) > 0$ for all x . By writing $\tilde{K}(x, y) = \pi(x)K(x, y)$, $\tilde{\mu}(x) = \mu(x)/\pi(x)$, we have

$$\sum_{x=1}^n \mu(x) \frac{(Gu)(x)}{u(x)} = \sum_{x=1}^n \frac{\tilde{\mu}(x)}{u(x)} [(\tilde{K}u)(x) + \frac{1}{2}(\Gamma u)(x)].$$

Hence it suffices to prove the result for $G = K + \frac{1}{2}\Gamma$, with K symmetric and Γ anti-symmetric.

Now take $u(x) = \sqrt{\mu(x)}$. Then

$$\begin{aligned} -\sum_{x=1}^n \mu(x) \frac{(Gu)(x)}{u(x)} &= -\sum_{x,y=1}^n \sqrt{\mu(x)} K(x, y) \sqrt{\mu(y)} \\ &\quad - \frac{1}{2} \sum_{x,y=1}^n \sqrt{\mu(x)} \Gamma(x, y) \sqrt{\mu(y)}. \end{aligned}$$

The first term is equal to the the rate function $I_K(\mu)$ for the empirical measure in the symmetric case, see e.g. (Hollander 2000, Theorem IV.14). Since Γ is skew-symmetric, the second term vanishes. Hence taking the supremum over u gives a value larger than or equal to $I_K(\mu)$.

Let u^* be given by $u^*(x) = \sqrt{\mu(x)/\pi(x)}$. Then

$$\begin{aligned} \nabla_u \left(-\sum_{x=1}^n \mu(x) \left(\frac{\frac{1}{2\pi(x)} \sum_{y=1}^n \Gamma(x, y) u(y) + \sum_{y=1}^n \gamma(x, y) K(x, y) u(y)}{u(x)} \right) \right) \Big|_{u^*} \\ = \Gamma u^*, \end{aligned}$$

which proves the second statement in the proposition.

If $\mu(x) = 0$ for some x , the proof carries over by only summing over indices for which $\mu(x) > 0$. \square

3.3 Mixing time and spectral gap

Non-reversibility in a Markov chain can have very favorable effects on mixing time and spectral gap, but we are not aware of general results in this direction. The reader is referred to Chen et al. (1999), Diaconis et al. (2000), Levin et al. (2009) for theoretical results in this direction, and to e.g. Sun et al. (2010), Turitsyn et al. (2011), Vucelja (2014) for less rigorous and/or numerical results. In these experimental results, non-reversibility especially seems to improve sampling in the case of sampling from a multimodal distribution.

4 NRMH in Euclidean space

In this section we explain how to extend the idea of non-reversible Metropolis-Hastings algorithm to a Euclidean state space. In Appendix 1 it is discussed how Metropolis-Hastings can be applied in the case of a general measurable space, and the discussion in this section is a special case of this.

4.1 General setting

Suppose Q is a Markov transition kernel on \mathbb{R}^n which has density $q(x, y)$ with respect to Lebesgue measure, i.e. $Q(x, rmdy) = q(x, y)dy$. We will be interested in sampling from a distribution on \mathbb{R}^n with Lebesgue density π . We do not require that $\int_{\mathbb{R}^n} \pi(x) dx = 1$. Let $\gamma : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be Lebesgue measurable and furthermore suppose γ satisfies

$$\gamma(x, y) = -\gamma(y, x) \quad \text{for all } x, y \in \mathbb{R}^n, \quad (11)$$

and

$$\int_{A \times \mathbb{R}^n} \gamma(x, y) dx dy = 0, \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^n). \quad (12)$$

Here $\mathcal{B}(\mathbb{R}^n)$ denotes the Borel σ -algebra generated by open sets in \mathbb{R}^n . Furthermore suppose that

$$\begin{aligned} \gamma(x, y) &= 0 \\ \text{for all } x, y \in \mathbb{R}^n \text{ for which } \pi(x)q(x, y) &= 0, \\ \pi(x)q(x, y) &= 0 \\ \text{if and only if } \pi(y)q(y, x) &= 0, \quad \text{for all } x, y \in \mathbb{R}^n, \end{aligned} \quad (13)$$

and

$$\begin{aligned} \gamma(x, y) + \pi(y)q(y, x) &\geq 0, \\ \text{for all } x, y \in \mathbb{R}^n \text{ for which } \pi(x)q(x, y) &\neq 0. \end{aligned} \quad (15)$$

Define the Hastings ratio

$$R(x, y) := \begin{cases} \frac{\gamma(x, y) + \pi(y)q(y, x)}{\pi(x)q(x, y)}, & \pi(x)q(x, y) \neq 0, \\ 1, & \pi(x)q(x, y) = 0, \end{cases} \quad (x, y \in \mathbb{R}^n), \quad (16)$$

acceptance probabilities $A(x, y) := \min(1, R(x, y))$, and transition kernel

$$\begin{aligned} P(x, B) &:= \int_B A(x, y) q(x, y) dy \\ &\quad + \left(1 - \int_S A(x, y) q(x, y) dy \right) \mathbb{1}_{x \in B}. \end{aligned}$$

The proof of the following theorem can be found in Appendix 1.

Theorem 4.1 *With the above notation and definitions, and assuming conditions (11), (12), (13), (14) and (15) are satisfied, P is a Markov transition kernel with invariant density π .*

In analogy with the discrete state space setting we call γ the *vorticity density* of (P, π) . If $\gamma \neq 0$ on a set of positive

Lebesgue measure then P is non-reversible, i.e. there exist sets $B_1, B_2 \subset \mathbb{R}^n$ such that

$$\int_{B_1} \left\{ \int_{B_2} \pi(x) P(x, dy) \right\} dx \neq \int_{B_2} \left\{ \int_{B_1} \pi(x) P(x, dy) \right\} dx.$$

4.2 Langevin diffusions for sampling in Euclidean space

The application of non-reversible sampling methods in Euclidean space is a relatively unexplored area. In this short review section we will discuss the use of Langevin diffusions for simulating from a target density, and discuss the potential role of NRMH in within this context.

Assume that the target density π is continuously differentiable. It is well known that π is invariant for the *Langevin diffusion*

$$dX(t) = \nabla(\log \pi)(X(t)) dt + \sqrt{2} dW(t), \quad t \geq 0. \quad (17)$$

where W is a standard Brownian motion in \mathbb{R}^n . A natural (discrete time) Markov chain for sampling π is then the *Euler-Maruyama discretization* of the Langevin diffusion,

$$X_{k+1} \sim N(X_k + h \nabla(\log \pi)(X_k), 2h), \quad k = 0, 1, 2, \dots \quad (18)$$

Here h is a suitable stepsize. This discretization is approximately correct if h is chosen to be sufficiently small. However, such a choice of h results in slow convergence to equilibrium of the Markov chain. When h is large, then the discretization results in large discrepancy between (17) and (18). As a result also the respective invariant distributions will be different and in particular the invariant distribution of (18) will not correspond to the desired distribution with density π . It is customary to correct for this by considering the Euler-Maruyama discretization as a proposal for Metropolis-Hastings, resulting in the Metropolis Adjusted Langevin Algorithm [MALA, Roberts and Tweedie (1996), Roberts and Rosenthal (1998)].

Any diffusion of the form

$$dX(t) = -(I + S) \nabla(\log \pi)(X(t)) dt + \sqrt{2} dW(t), \quad t \geq 0, \quad (19)$$

with $S \in \mathbb{R}^{n \times n}$ skew-symmetric, has π as invariant density. If $S \neq 0$ then the diffusion is non-reversible. In analogy with Sect. 3 one may hope that such a non-reversible diffusion has advantages compared to the reversible Langevin diffusion. In fact for multivariate Gaussian target distributions these advantages are clear Hwang et al. (1993), Lelièvre

et al. (2013), as we will discuss below. More generally¹ the probability of large deviations of the empirical distribution from the invariant distribution are reduced for non-reversible diffusions Rey-Bellet and Spiliopoulos (2014).

When discretizing (19) it is again necessary to correct for discretization error by a MH accept/reject step. However, since MH generates reversible chains, one should expect that also the favourable properties of non-reversibility are destroyed. Instead, an implementation of NRMH should be able to preserve these favourable properties of non-reversible diffusions. We will illustrate this for multivariate Gaussian distributions.

4.3 Non-reversible Metropolis-Hastings for sampling multivariate Gaussian distributions

Consider as target distribution a centered normal distribution with positive definite covariance matrix V . In this case the Langevin diffusion becomes the Ornstein-Uhlenbeck process

$$dX(t) = -V^{-1} X(t) dt + \sqrt{2} dW(t), \quad (20)$$

where $(W(t))$ is an n -dimensional standard Brownian motion. In Hwang et al. (1993), it is shown that adding a ‘nonreversible’ component of the form $-SV^{-1}$ to the drift, with S skew-symmetric, can improve convergence of the sample covariance. Therefore we will instead consider the Ornstein-Uhlenbeck process with modified drift

$$dX(t) = B X(t) dt + \sqrt{2} dW(t), \quad (21)$$

where $B := -(I + S)V^{-1}$ with S skew-symmetric. For any choice of skew-symmetric S , this diffusion keeps π invariant. The convergence to equilibrium of the diffusion is governed by the spectral bound, $\mathfrak{s}(B) := \max\{\operatorname{Re} \lambda : \lambda \in \sigma(B)\}$. More specifically,

$$\begin{aligned} \operatorname{Cov}(X(t)) &= 2 \int_0^t e^{Bs} e^{B's} ds \\ &\rightarrow 2 \int_0^\infty e^{Bs} e^{B's} ds = V \quad (\text{as } t \rightarrow \infty), \end{aligned}$$

with rate of convergence

$$\frac{1}{t} \ln \left\| \int_t^\infty e^{Bs} e^{B's} ds \right\| \rightarrow 2\mathfrak{s}(B).$$

Also, $\mathfrak{s}(B) \leq \mathfrak{s}(-V^{-1})$ for any choice of S . In other words, adding a non-reversible term increases the speed of convergence to equilibrium. In Lelièvre et al. (2013), it is

¹ Strictly speaking, the analysis of Rey-Bellet and Spiliopoulos (2014) applies to diffusions on a compact manifold.

established that it is possible to choose S optimally, resulting in $\mathfrak{s}(B) = -\text{tr}(V^{-1})/n$. By choosing S in such a way the convergence of the chain is effectively governed by the average of the eigenvalues, which should be compared to the reversible case in which the ‘worst’ eigenvalue determines the speed of convergence.

We will apply the theory developed in Sect. 4.1 to the time discretization of the non-reversible Ornstein-Uhlenbeck process. To be able to satisfy (15) later on, we will require flexibility in the magnitude of the drift multiplier B and the diffusivity. We consider the time discretization of (21), with step size $h > 0$, is

$$X_{k+1} = X_k + hBX_k + \sqrt{2h\sigma}Z_{k+1}, \quad (22)$$

where (Z_k) are i.i.d. standard normal and $\sigma > 0$. For $\sigma = 1$ this is the usual Euler-Maruyama discretization. The transition kernel of the Euler-Maruyama discretization will serve as our proposal distribution $Q(x, dy) = q(x, y)dy$, and we will first determine the vorticity density γ of Q .

Let $\mathfrak{r}(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$ denote the spectral radius of a square matrix A . Provided $\mathfrak{r}(I + hB) < 1$, the invariant probability distribution of Q is the centered normal distribution with covariance R , where R is the unique positive definite matrix solution to the *discrete time Lyapunov equation* [see e.g. (Lancaster and Tismenetsky 1985, Theorem 13.2.1)]

$$R = 2h\sigma^2 I + (I + hB)R(I + hB'). \quad (23)$$

Let ρ denote the density of the invariant probability distribution of Q . Let $f(x, y) = \rho(x)q(x, y)$. The vorticity density of the proposal chain is

$$\gamma(x, y) := f(x, y) - f(y, x). \quad (24)$$

As target density we have

$$\pi(x) = ((2\pi)^n \det V)^{-1/2} \exp\left(-\frac{1}{2}x'V^{-1}x\right), \quad x \in \mathbb{R}^n.$$

It is clear that γ satisfies (11) and (12). Also since π and q are non-degenerate, (13) and (14) are satisfied. The same statements hold trivially for scalar multiples of γ .

Verification of (15) requires more effort. We provide a sufficient condition. The proof of this result is provided in Appendix 1.

Theorem 4.2 Define constants $0 < C_1 \leq C_2$ by

$$\begin{aligned} C_1 &= \|V^{-1/2}(I + S)V^{-1}(I - S)V^{1/2}\| \quad \text{and} \\ C_2 &= \|V^{-1/2}(I + S)V^{-1/2}\|^2 \|V\|. \end{aligned} \quad (25)$$

Suppose $c > 0$, $h > 0$ and $\sigma > 0$ satisfy

$$h < \frac{2}{C_2}, \quad \sigma^2 \leq \frac{2 - hC_2}{2 - h(C_2 - C_1)}, \quad \text{and} \quad c \leq \sigma^n. \quad (26)$$

Then $\tilde{\gamma}(x, y) := c\gamma(x, y)$, with γ as constructed above, satisfies (15), and is therefore a vorticity density compatible with proposal distribution $Q(x, dy) \sim N((I + hB)x, 2h\sigma^2 I)$ and invariant distribution $N(0, V)$.

Remark 4.3 How should one choose c , h and σ ? It seems reasonable to choose σ^2 equal to the maximal allowed value in (26), so that the deviation from the Euler-Maruyama discretization (which has $\sigma = 1$) is minimal; i.e. let

$$\sigma = \sigma(h) = \sqrt{\frac{2 - hC_2}{2 - h(C_2 - C_1)}}. \quad (27)$$

To maximize the non-reversibility effects in the acceptance probability one should choose c as large as possible, i.e. $c = \sigma^n(h)$. A heuristic estimate for the scaling of the expected step size is given by the step size h times the multiplicative factor in the vorticity, $c = \sigma^n(h)$ in the acceptance probability. Note that $h\sigma^n(h) = 0$ for $h = 0$ and $h = \frac{2}{C_2}$. Maximization of $h\sigma^n(h)$ with respect to h yields

$$h = \frac{2}{C_2} + \frac{(n+2)C_1}{2C_2(C_2 - C_1)} - \frac{\sqrt{(n-2)^2 C_1^2 + 8n C_1 C_2}}{2C_2(C_2 - C_1)} \quad (28)$$

as long as $C_1 < C_2$ (which is the case in which S and V do not commute). This expression satisfies the condition $h < \frac{2}{C_2}$. A first order Taylor approximation around $1/n$ yields the simplified expression $h \approx \frac{4}{C_2(n+2)} < \frac{2}{C_2}$. The corresponding value of $\sigma^2(h)$ is to first order equal to $\sigma^2(h) \approx 1 - \frac{2C_1}{C_2(n+2)}$.

In case $C_1 = C_2$, the optimal value of h is given by $h = \frac{4}{(n+2)C_2}$, with $\sigma^2(h) = 1 - \frac{2}{n+2}$.

To summarize, a general procedure for applying non-reversible Metropolis-Hastings may be described by the steps listed in Fig. 1.

4.4 Numerical experiments

Below we carry out two experiments illustrating the approach above. For the obtained Markov chain realization (X_1, \dots, X_P) , we will obtain an estimate of the decorrelation by considering the *empirical autocorrelation function (EACF)* r defined by

$$r^i(k) = \frac{1}{P-k} \sum_{p=1}^{P-k} (X_p^i - \hat{\mu}^i) (X_{p+k}^i - \hat{\mu}^i),$$

Fig. 1 Non-reversible Metropolis-Hastings with Ornstein-Uhlenbeck proposals

Algorithm for sampling from a Gaussian multivariate distribution

Given:

- target distribution $N(0, V)$, with V positive definite.

Initialization:

- (1) Determine a skew-symmetric $S \in \mathbb{R}^{n \times n}$ such that the magnitude of $\mathfrak{s}(-(I + S)V^{-1})$ is large, e.g. using the algorithm of [LNP13], and let $B := -(I + S)V^{-1}$.
- (2) Choose c , h and σ satisfying (26), e.g. using the choices discussed in Remark 4.3.
- (3) Solve the discrete Lyapunov equation (23) to determine the invariant covariance matrix R of the discretized scheme.
- (4) Let M , f and γ be given by (40), (41) and (24).
- (5) Let $\tilde{\gamma}(x, y) := c\gamma(x, y)$ for $x, y \in \mathbb{R}^n$.
- (6) Choose a starting point $X_0 \in \mathbb{R}^n$.

Sampling: Repeat for $k = 1, 2, \dots$

- (1) Generate proposal $Y \sim N((I + hB)X_{k-1}, 2h\sigma^2 I)$
- (2) Compute $R(X_{k-1}, Y)$ according to (16), where $q(x, \cdot)$ is taken to be equal to the density function of the proposal distribution $N((I + hB)x, 2h\sigma^2 I)$
- (3) With probability $1 \wedge R(X_k, Y)$, let $X_k = Y$; otherwise let $X_k = X_{k-1}$.

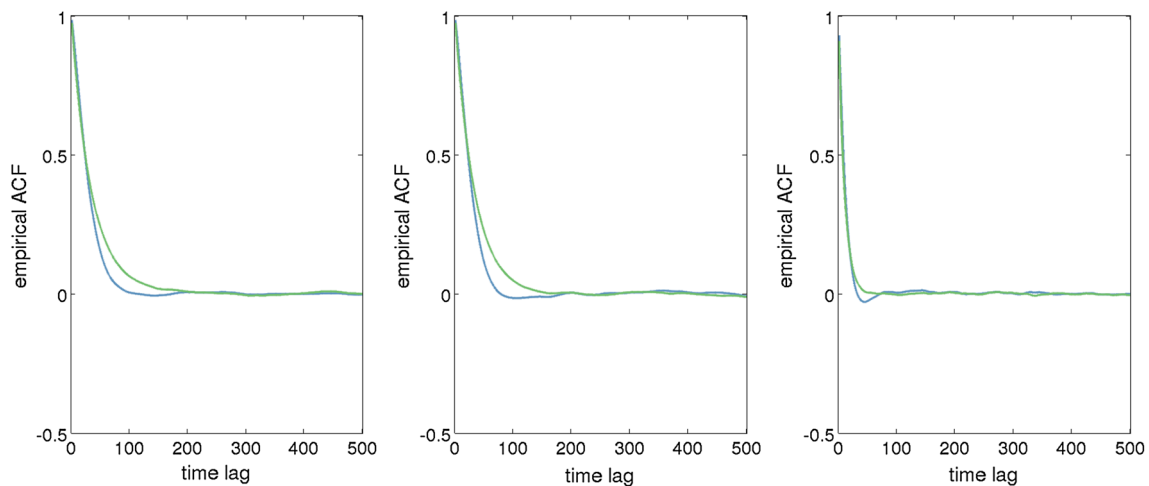


Fig. 2 Empirical autocorrelation functions for 3-dimensional example of Sect. 4.4.1. The blue curve represents the non-reversible Metropolis-Hastings method. The green curve represents reversible proposals in

conjunction with the usual MH acceptance rule. The step sizes h for all approaches are taken to be the same. These plots are based on an MCMC trajectory of 10^6 steps. (Color figure online)

where i ranges over the coordinates $i = 1, \dots, n$, and where $\hat{\mu}^i = \frac{1}{P} \sum_{p=1}^P X_p^i$ is the empirical average of the i -th coordinate. A fast decaying EACF indicates that the samples generated by the Markov chain are quickly decorrelating.

4.4.1 Three-dimensional example

In this example, from Hwang et al. (1993), we take as target covariance structure V a diagonal matrix with $\text{diag}(V) = (1, 1, 1/4)$ s. The optimal nonlinear drift is obtained by letting

$$S = \begin{pmatrix} 0 & \sqrt{3} & 1 \\ -\sqrt{3} & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix}$$

We choose the parameter values in accordance with Remark 4.3, resulting in

$$c = 0.5333, \quad h = 0.0334, \quad \sigma = 0.8109.$$

The performance of NRMH is compared to MH with identical step-size h , and reversible proposal distribution $Q(x, dy) \sim N((I - hV^{-1})x, 2hI)$. In Fig. 2 the EACFs for this 3-dimensional example are plotted. Here we see that NRMH helps to decrease the autocorrelations of the slowly decorrelating components in MH (here, the first two components). It achieves this without increasing autocorrelations of components that are already quickly decorrelating (here, the third component).

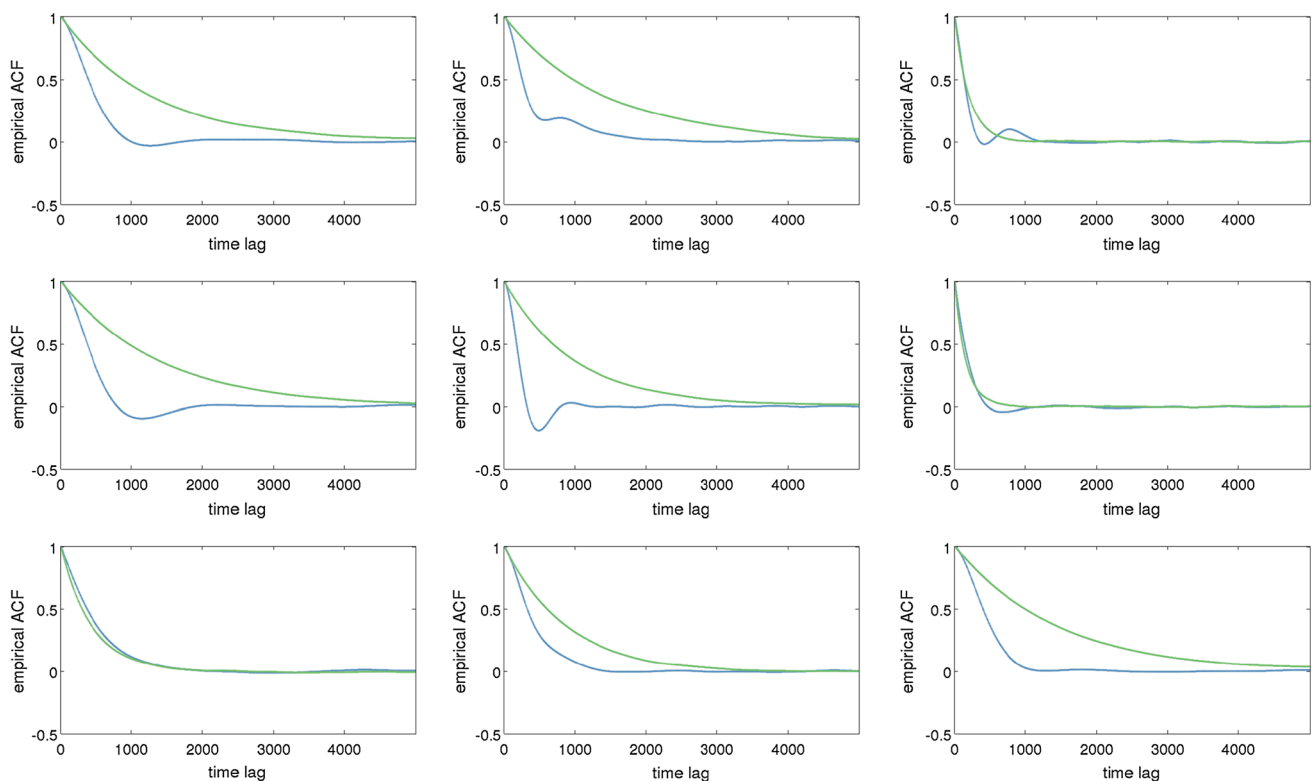


Fig. 3 Empirical autocorrelation functions for the 9-dimensional example of Sect. 4.4.2. The *blue curve* represents the non-reversible Metropolis Hastings method. The *green curve* represents reversible pro-

posals in conjunction with the usual MH acceptance rule. These plots are based on an MCMC trajectory of 10^7 steps. (Color figure online)

4.4.2 Nine-dimensional example

Here we generated a random diagonal matrix V , with

$$\text{diag}(V) = (0.8147, 0.9058, 0.1270, 0.9134, 0.6324, 0.0975, 0.2785, 0.5469, 0.9575).$$

Using the algorithm described in Lelièvre et al. (2013) an optimal non-reversible drift $B = -(I + S)V^{-1}$ can be computed. For reversible dynamics, we have $\mathfrak{s}(-V^{-1}) = -1.0444$, while for the optimal non-reversible dynamics, $\mathfrak{s}(B) = -\text{tr } V^{-1}/n = -3.2891$. In Fig. 3 the EACFs for this 9-dimensional example are plotted. One can clearly see the typical effect of adding non-reversibility: the autocorrelation of the worst coordinates is improved so that it becomes on par with that of the fastest decorrelating coordinates.

In this case, choosing c , h and σ as in Remark 4.3 results in values

$$c = 0.4313, \quad h = 7.0822 \times 10^{-4} \quad \text{and} \quad \sigma = 0.9108.$$

In a numerical example with 10^7 proposed transitions this leads to acceptance ratios displayed in Table 2. Using the batch means method (by dividing the sample trajectory in

\sqrt{n} trajectories of length \sqrt{n} and assuming the \sqrt{n} are independent), we can estimate asymptotic variance. The resulting estimates for asymptotic variance of the different components are given in Table 1. It should be noted that the notion of asymptotic variance is only defined in case a CLT holds [see Roberts and Rosenthal (2004)], which strictly speaking is an open question in this setting.

It is well known that for optimal convergence in Metropolis Adjusted Langevin (MALA), the stepsize should be tuned so that the ratio of accepted proposals is approximately equal to 0.574 Roberts and Rosenthal (1998). Compared to this, the acceptance ratios in our example, given in Table 2, are fairly high. In particular the Metropolis-Hastings chain can be improved significantly by increasing step size h , and so we are currently comparing NRMH with a sub-optimal tuning of MH.

The results of this experimental section should therefore be considered as a proof of concept of NRMH in continuous state spaces, rather than as an advertisement for its immediate practicality. The experiments do illustrate the faster decorrelation of NRMH in comparison to MH (for fixed step-size). It is an open question if the framework of NRMH can be extended so that NRMH would become competitive with optimally tuned MH.

Table 1 Estimated asymptotic variances for the 9-dimensional example of Sect. 4.4.2, based on an MCMC trajectory of 10^7 steps

Component	1	2	3	4	5	6	7	8	9
NRMH	599.96	661.17	40.80	572.26	159.05	27.35	230.41	401.98	718.64
MH	1315.3	1522.2	47.156	1473.3	876.46	28.316	204.05	708.83	1578.2

Note that NRMH improves the (estimated) asymptotic variance for almost all components (and often significantly), except for component #7

Table 2 Acceptance ratios for 9-dimensional example of Sect. 4.4.2, for a sample path consisting of 10^7 proposals, and with a step size $h = 7.0822 \times 10^{-4}$

	Acceptance ratio
NRMH	0.7383
MH	0.9343

5 Discussion

The efficiency increase of non-reversible Markov chains in MCMC can be significant, in terms of either asymptotic variance or mixing properties, as remarked in this paper. NRMH extends the MCMC-toolbox with a method to utilize these benefits. In particular for continuous state spaces it was, to our knowledge, not known how to construct non-reversible Markov chains for MCMC sampling (taking into account the necessity of a correction step when using time discretization of diffusions).

Using the theory developed in Sect. 4, NRMH can be applied to general distributions on \mathbb{R}^n as follows. For a target density function π , suppose there exists a Gaussian distribution $N(0, V)$ with density function π_0 on \mathbb{R}^n , satisfying $k\pi_0 \leq \pi$ on \mathbb{R}^n for some $k > 0$. Then if γ is a suitable vorticity density for sampling from $N(0, V)$, using proposal density $q(x, y)$, we have for $\tilde{\gamma} := k\gamma$ that

$$\begin{aligned}\tilde{\gamma}(x, y) + \pi(x)q(x, y) &= k\gamma(x, y) + \pi(x)q(x, y) \\ &\geq k(\gamma(x, y) + \pi_0(x)q(x, y)) \geq 0,\end{aligned}$$

so that (15) is satisfied for the combination for this choice of π , $\tilde{\gamma}$ and q , and Theorem 4.1 applies. Such a suitable choice of γ may be determined as described in Sect. 4.3.

The approach outlined in Sect. 4 should be considered as a first attempt at implementing the NRMH framework for continuous spaces. As discussed, in order to use the framework one needs to verify the non-negativity condition which leads to technical challenges. In particular we expect that much progress is possible in weakening the conditions of Theorem 4.2. The current form of that proposition results in a relatively small step size h , which obstructs fast convergence of NRMH. As mentioned before, it is an open question whether NRMH in continuous spaces can be made competitive with optimally tuned MALA.

The theoretical discussion of Sect. 3 and the numerical experiment of Sect. 4.4 illustrate how efficiency can be improved by employing non-reversible Metropolis-Hastings. In view of these encouraging results it will hopefully be possible to extend the result to more general settings. The practical application of NRMH depends on the identification of suitable vorticity structures that are compatible with proposal chains, and establishing these in practical examples provides a promising and challenging direction of research.

Analysis of non-reversible Markov chains is difficult, essentially because self-adjointness is lost. Without self-adjointness, it is much more difficult to connect spectral theory to mixing properties of chains. It seems that a good way of understanding benefits of non-reversible sampling is by studying Cesaro averages [see Levin et al. (2009) and e.g. the result on large deviations in Sect. 3.2]. The results of Sect. 3 which establish that non-reversible chains have better asymptotic variance or large deviations properties, are so far qualitative in nature (i.e. fail to quantify the amount of improvement). To obtain quantitative results is an important challenge that remains to be addressed. Also, it is object of further study how these results carry over to countable and uncountable state spaces. In particular, the question under what conditions the resulting chains are geometrically ergodic and/or satisfy a CLT should be considered.

Acknowledgments I am grateful to Prof. Pavliotis (Imperial College, London) for making available the code for computing optimal non-reversible drift (as discussed in Lelièvre et al. (2013)). I also wish to acknowledge valuable discussions with Prof. Hilbert J. Kappen (Radboud University), Dr. Kevin Sharp (University of Oxford) and Prof. Gareth Roberts (University of Warwick). We thank the reviewers and editor for their valuable suggestions which have had a significant impact upon the paper. This research has received support from the European Union project # FP7-ICT-270327 (Complacs), and the EPSRC under the CRiSM Grant: EP/D002060/1.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix 1: NRMH in general state spaces

Let (S, \mathcal{S}) be a measurable space. Let $P(x, dy)$ denote a Markov transition kernel and π an invariant probability distribution of P , i.e. $\int_S P(x, A)\pi(dx) = \pi(A)$ for $A \in$

\mathcal{S} . Define $F_P(dx, dy) := \pi(dx)P(x, dy)$, $B_P(dx, dy) = \pi(dy)P(y, dx)$. Here F and B in F_P and B_P denote ‘forward’ and ‘backward’, respectively. Note that F_P and B_P are probability measures on $S \times S$ with marginal distributions π .

Define

$$\begin{aligned}\Gamma(dx, dy) &:= F_P(dx, dy) - B_P(dx, dy) \\ &= \pi(dx)P(x, dy) - \pi(dy)P(y, dx).\end{aligned}\quad (29)$$

Then Γ is a signed measure on $S \times S$, satisfying

$$\Gamma(A \times B) = -\Gamma(B \times A), \quad A, B \in \mathcal{S}, \quad (30)$$

and

$$\Gamma(A, S) = 0, \quad A \in \mathcal{S}. \quad (31)$$

We will call a signed measure Γ on $S \times S$ satisfying (30), (31) a *vorticity measure*. If Γ is related to π and P by (29), it is called the *vorticity measure of* (P, π) .

Let $Q(x, dy)$ and let F_Q and B_Q as defined above with P replaced by Q . Let Γ be a vorticity measure. The Markov chain Q will play the role of proposal chain, and Γ the role of target vorticity.

Definition 5.1 (*Absolute continuity of Γ*) We can use the Jordan decomposition (Rudin 1987, Sect. 6.6) to decompose Γ into two non-signed measures $\Gamma^+ := \frac{1}{2}(|\Gamma| + \Gamma)$ and $\Gamma^- := \frac{1}{2}(|\Gamma| - \Gamma)$, so that $\Gamma = \Gamma^+ - \Gamma^-$. We say that Γ is *absolutely continuous* with respect to some measure M on $S \times S$, denoted by $\Gamma \ll M$, if $\Gamma^- \ll M$ and $\Gamma^+ \ll M$. If $\Gamma \ll M$, we define the Radon-Nikodym derivative of Γ with respect to M by

$$\frac{d\Gamma}{dM}(x, y) = \frac{d\Gamma^+}{dM}(x, y) - \frac{d\Gamma^-}{dM}(x, y), \quad x, y \in S.$$

Assuming $B_Q \ll F_Q$ and $\Gamma \ll F_Q$, we define the *non-reversible Hastings ratio*

$$R(x, y) := \frac{d\Gamma}{dF_Q}(x, y) + \frac{dB_Q}{dF_Q}(x, y). \quad (32)$$

In order for $R(x, y)$ to be nonnegative, we have to impose the condition

$$\frac{d\Gamma}{dF_Q}(x, y) + \frac{dB_Q}{dF_Q}(x, y) \geq 0, \quad x, y \in S. \quad (33)$$

Lemma 5.2 Suppose Q and π are such that F_Q and B_Q are equivalent measures on $S \times S$. Suppose that Γ is a vorticity measure such that Γ is absolutely continuous with respect to F_Q . Suppose (33) is satisfied. Then for all $x, y \in S$, $R(y, x) \geq 1$ if and only if $R(x, y) \leq 1$.

Proof Write $\Gamma = \Gamma^+ - \Gamma^-$ for the Jordan decomposition of Γ , i.e. Since $\Gamma(A, B) = -\Gamma(B, A)$, it follows that $|\Gamma|(A, B) = |\Gamma|(B, A)$ and $\Gamma^+(A, B) = \Gamma^-(B, A)$ for $A, B \in \mathcal{S}$. Suppose for $x, y \in S$, $R(x, y) \leq 1$, so that

$$\begin{aligned}1 \geq \frac{d(\Gamma + B_Q)}{dF_Q}(x, y) &= \frac{d\Gamma^+}{dF_Q}(x, y) - \frac{d\Gamma^-}{dF_Q}(x, y) \\ &\quad + \frac{dB_Q}{dF_Q}(x, y).\end{aligned}\quad (34)$$

We compute

$$\begin{aligned}R(y, x) &= \frac{d(\Gamma + B_Q)}{dF_Q}(y, x) = \frac{d\Gamma^+}{dF_Q}(y, x) - \frac{d\Gamma^-}{dF_Q}(y, x) \\ &\quad + \frac{dB_Q}{dF_Q}(y, x) \\ &= \frac{d\Gamma^-}{dF_Q}(x, y) - \frac{d\Gamma^+}{dF_Q}(x, y) + \frac{dF_Q}{dB_Q}(x, y) \\ &= \frac{dF_Q}{dB_Q}(x, y) \left(\frac{d\Gamma^-}{dF_Q}(x, y) - \frac{d\Gamma^+}{dF_Q}(x, y) + 1 \right) \\ &\geq \frac{dF_Q}{dB_Q}(x, y) \frac{dB_Q}{dF_Q}(x, y) = 1,\end{aligned}$$

where (34) was used to establish the inequality. \square

Define the *acceptance probability* by

$$A(x, y) = \min(1, R(x, y)), \quad x, y \in S, \quad (35)$$

and define a transition kernel P by

$$\begin{aligned}P(x, dy) &= A(x, y)Q(x, dy) \\ &\quad + \left(1 - \int_S A(x, z)Q(x, dz)\right) \delta_x(dy).\end{aligned}\quad (36)$$

Lemma 5.3 Suppose Q and π are such that F_Q and B_Q are equivalent measures on $S \times S$. Suppose that Γ is a vorticity measure such that Γ is absolutely continuous with respect to F_Q . Suppose (33) is satisfied. Let P be defined by (35), (36). The vorticity measure of (P, π) is Γ .

Proof We should check that for $A, B \in \mathcal{S}$,

$$\begin{aligned}\Gamma(A, B) &= \int_A \left(\int_B P(x, dy) \right) \pi(dx) \\ &\quad - \int_B \left(\int_A P(y, dx) \right) \pi(dy).\end{aligned}\quad (37)$$

Note that, for some measurable function $h : S \rightarrow \mathbb{R}$,

$$\begin{aligned}\int_A \left(\int_B h(x) \delta_x(dy) \right) \pi(dx) &= \int_A h(x) \mathbb{1}_{x \in B} \pi(dx) \\ &= \int_{A \cap B} h(x) \pi(dx).\end{aligned}$$

This shows that the measure $(1 - \int_S A(x, z)Q(x, dz))\delta_x(dy)\pi(dx)$ in (36) is symmetric with respect to interchanging x and y , and therefore has no contribution in the verification of (37).

Using Lemma 5.2 there are no conceptual difficulties left in verifying that

$$\int_A \left(\int_B P(x, dy) \right) \pi(dx) - \int_B \left(\int_A P(y, dx) \right) \pi(dy) = \Gamma(A, B),$$

in a similar way as in the proof of Lemma 2.4. \square

Theorem 5.4 Suppose Q and π are such that F_Q and B_Q are equivalent measures on $S \times S$. Suppose that Γ is a vorticity measure such that Γ is absolutely continuous with respect to F_Q . Suppose (33) is satisfied. Let P be defined by (35), (36). Then π is invariant for P and Γ is the vorticity measure of (P, π) .

Proof Lemma 5.3 establishes that Γ is the vorticity measure of (P, π) . By 31, for $A \in S$,

$$0 = \Gamma(A, S) = \pi(A) - \int_S P(y, A)\pi(dy),$$

so that π is invariant for P . \square

Proof of Theorem 4.1 Define the signed measure Γ , and measures F_Q and B_Q on $S \times S$ by

$$\begin{aligned} \Gamma(dx, dy) &= \gamma(x, y) dx dy, \\ F_Q(dx, dy) &= \pi(dx)Q(x, dy), \\ B_Q(dx, dy) &= \pi(dy)Q(y, dx). \end{aligned}$$

Then by (11), (12), Γ is a vorticity measure, and by assumptions (13) and (14), F_Q and B_Q are equivalent measures, and Γ is absolutely continuous with respect to F_Q . Furthermore $R(x, y)$ is a version of $\frac{d\Gamma}{dF_Q} + \frac{dB_Q}{dF_Q}$, and by assumption (15), we have $\frac{d\Gamma}{dF_Q} + \frac{dB_Q}{dF_Q} \geq 0$. We see that all conditions of Theorem 5.4 are satisfied, so that the stated results follow. \square

Appendix 2: Proof of Theorem 4.2

Define an inner product $\langle \cdot, \cdot \rangle_V$ on \mathbb{R}^n by $\langle x, y \rangle_V = \langle x, V^{-1}y \rangle$, where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. Let $\| \cdot \|_V$ denote the induced norm. Let $B := -(I + S)V^{-1}$.

Lemma 5.5 Suppose

$$0 < h < \frac{2}{\|V^{-1/2}(I - S^2)V^{-1/2}\|}. \quad (38)$$

Then $\tau(I + hB) < 1$.

Proof Suppose $\lambda \in \sigma(I + hB)$ and let $x \in \mathbb{C}^n$, $x \neq 0$, be an eigenvector corresponding to λ . Without loss of generality assume that $\|x\|_V = \|V^{-1/2}x\| = 1$. We have

$$\lambda = \langle (I + hB)x, x \rangle_V = \|x\|_V^2 - h \langle (I + S)V^{-1}x, V^{-1}x \rangle.$$

By skew-symmetry of S ,

$$\operatorname{Re} \lambda = \|V^{-1/2}x\|^2 - h\|V^{-1}x\|^2 = 1 - h\|V^{-1}x\|^2,$$

and

$$|\operatorname{Im} \lambda| = h|\langle SV^{-1}x, V^{-1}x \rangle|.$$

Hence

$$\begin{aligned} |\lambda|^2 &= (\|V^{-1/2}x\|^2 - h\|V^{-1}x\|^2)^2 + h^2 \langle SV^{-1}x, V^{-1}x \rangle^2, \\ &= 1 - 2h\|V^{-1}x\|^2 + h^2 (\|V^{-1}x\|^4 + \langle SV^{-1}x, V^{-1}x \rangle^2). \end{aligned}$$

Hence the requirement $|\lambda| < 1$ translates into the inequality

$$h^2 (\|V^{-1}x\|^4 + \langle SV^{-1}x, V^{-1}x \rangle^2) < 2h\|V^{-1}x\|^2,$$

or, after some rearranging,

$$h < \frac{2}{\left(\|V^{-1}x\|^2 + \frac{\langle SV^{-1}x, V^{-1}x \rangle^2}{\|V^{-1}x\|^2} \right)}. \quad (39)$$

Looking at the denominator, using Cauchy-Schwartz,

$$\begin{aligned} \|V^{-1}x\|^2 + \frac{\langle SV^{-1}x, V^{-1}x \rangle^2}{\|V^{-1}x\|^2} &\leq \|V^{-1}x\|^2 + \|SV^{-1}x\|^2 = \langle V^{-1}x, V^{-1}x \rangle \\ &\quad + \langle SV^{-1}x, SV^{-1}x \rangle \\ &= \langle V^{-1}x, (I - S^2)V^{-1}x \rangle \\ &= \langle V^{-1/2}x, V^{-1/2}(I - S^2)V^{-1/2}V^{-1/2}x \rangle \\ &\leq \|V^{-1/2}(I - S^2)V^{-1/2}\| \|V^{-1/2}x\|^2 \\ &= \|V^{-1/2}(I - S^2)V^{-1/2}\|. \end{aligned}$$

It follows that if h satisfies (38), then it satisfies (39), and therefore $\tau(I + hB) < 1$. \square

If (38) holds, by (Lancaster and Tismenetsky 1985, Theorem 13.2.1) there exists a unique solution $R = R(\sigma)$ to the discrete Lyapunov Eq. (23). Recall $f(x, y) = \rho(x)q(x, y)$, where ρ is the density of $N(0, R)$ and $q(x, \cdot)$ the density function of $N((I + hB)x, (2h\sigma^2)I)$. Hence f is a Gaussian density function with mean zero and covariance matrix

$$M := \begin{pmatrix} R & R(I + hB') \\ (I + hB)R & 2h\sigma^2 I + (I + hB)R(I + hB') \end{pmatrix} \quad \text{(iii) For } h < \frac{2}{C_2} \text{ we have}$$

$$= \begin{pmatrix} R & R(I + hB') \\ (I + hB)R & R \end{pmatrix}. \quad (40) \quad \|BRB'\| \leq \frac{2\sigma^2 C_1}{2 - hC_2}.$$

Lemma 5.6 Suppose (38) holds. Then $\det M = (2h\sigma^2)^n \det(R)$

Proof By standard result on determinants of block matrices,

$$\begin{aligned} \det M &= \det \begin{pmatrix} R & R(I + hB') \\ (I + hB)R & R \end{pmatrix} \\ &= \det(R) \det \left(R - (I + hB)RR^{-1}R(I + hB') \right) \\ &= \det(R) \det \left(R - (I + hB)R(I + hB') \right). \end{aligned}$$

In the argument of the second determinant we recognize (23), from which we obtain

$$\det M = \det(R) \det(2h\sigma^2 I_n) = (2h\sigma^2)^n \det(R). \quad \square$$

Let \leq denote the partial ordering of positive definite matrices, i.e. $A \geq B$ if $A - B$ is positive semidefinite.

Lemma 5.7 Suppose (38) holds. Then $R \geq \sigma^2 V$.

Proof of Theorem 4.1 Expanding (23) gives

$$R = 2h\sigma^2 I + R + hBR + hRB' + h^2BRB',$$

or equivalently

$$\begin{aligned} 0 &= 2\sigma^2 I + (BR + RB') + hBRB' = 2\sigma^2 I \\ &\quad + (BR + RB') + T, \end{aligned}$$

where $T := hBRB' \geq 0$. It follows that R satisfies the continuous time Lyapunov equation

$$BR + RB' = -(2\sigma^2 I + T)$$

with solution [see e.g. (Lancaster and Tismenetsky 1985, Theorem 13.1.1)]

$$\begin{aligned} R &= \int_0^\infty e^{Bs} (2\sigma^2 I + T) e^{B's} ds \\ &\geq \int_0^\infty e^{Bs} (2\sigma^2 I) e^{B's} ds = \sigma^2 V, \end{aligned}$$

where the last equality follows because V satisfies the continuous time Lyapunov equation $BV + VB' = -2I$. \square

Lemma 5.8 (i) For all $x \in \mathbb{R}^n$, $\langle x, Bx \rangle_V \leq -\|x\|_V^2 / \|V\|$.

(ii) For all $s \geq 0$, $\|e^{Bs}\|_V \leq e^{-s/\|V\|}$.

Define constants $0 \leq C_1 < C_2$ by (25).

(iv) For $h < \frac{2}{C_2}$, we have

$$R \leq \sigma^2 \left(\frac{2 - h(C_2 - C_1)}{2 - hC_2} \right) V.$$

Proof Denote $\omega := \min\{\lambda : \lambda \in \sigma(V^{-1})\} = 1/\|V\|$.

(i) Since S is skew-symmetric,

$$\begin{aligned} \langle x, Bx \rangle_V &= -\langle V^{-1}x, (I + S)V^{-1}x \rangle \\ &= -\langle V^{-1}x, V^{-1}x \rangle = -\|V^{-1/2}y\|^2, \end{aligned}$$

where $y = V^{-1/2}x$. Now $\|V^{-1/2}y\| \geq (1/\|V\|^{1/2})\|y\|$, so we conclude

$$\langle x, Bx \rangle_V \leq -\|y\|^2 / \|V\| = -\|x\|_V^2 / \|V\|.$$

(ii) This follows immediately from (i) and the fact that a dissipative operator (here: $B + \frac{1}{\|V\|}I$) generates a contraction semigroup (Yosida 1980, Sect. IX.8).

(iii) First note that

$$\begin{aligned} &\|V^{-1/2}(I - S^2)V^{-1/2}\| \\ &= \|V^{-1/2}(I + S)V^{-1/2}V^{-1/2}(I - S)V^{-1/2}\| \leq C_2. \end{aligned}$$

Therefore if $h < \frac{2}{C_2}$, then (38) holds, and therefore R is well-defined. From the proof of Lemma 5.7, $R = \sigma^2 V + h \int_0^\infty e^{Bs} BRB' e^{B's} ds$. Hence, using (ii),

$$\begin{aligned} &\|BRB'\|_V \\ &= \left\| B \left(\sigma^2 V + h \int_0^\infty e^{Bs} BRB' e^{B's} ds \right) B' \right\|_V \\ &\leq \sigma^2 \|BV B'\|_V + h \|B\|_V^2 \left(\int_0^\infty \|e^{Bs}\|_V^2 ds \right) \|BRB'\|_V \\ &\leq \sigma^2 \|BV B'\|_V + \frac{1}{2} h \|B\|_V^2 \|V\| \|BRB'\|_V. \end{aligned}$$

The result follows after rearranging, using the following equality (which holds for any matrix K) to express $\|B\|_V$ in terms of the $\|\cdot\|$ -norm:

$$\begin{aligned} \|K\|_V &= \sup_{x \neq 0} \frac{\|V^{-1/2}Kx\|}{\|V^{-1/2}x\|} = \sup_{y \neq 0} \frac{\|V^{-1/2}KV^{1/2}y\|}{\|y\|} \\ &= \|V^{-1/2}KV^{1/2}\|. \end{aligned}$$

(iv) By (iii), using that $\tau(BRB') \leq |||BRB' |||$ for any matrix-norm $||| \cdot |||$,

$$BRB' \leq \left(\frac{2\sigma^2 C_1}{2 - hC_2} \right) I.$$

Hence

$$\begin{aligned} R &= \sigma^2 V + h \int_0^\infty e^{Bs} BRB' e^{B's} ds \leq \sigma^2 V \\ &\quad + \left(\frac{2h\sigma^2 C_1}{2 - hC_2} \right) \int_0^\infty e^{Bs} e^{B's} ds \\ &= \sigma^2 \left(1 + \frac{hC_1}{2 - hC_2} \right) V, \end{aligned}$$

which is equivalent to the stated result. \square

Proof The density f is given by

$$f(x, y) = (2\pi)^{-n} (\det M)^{-1/2} \exp \left(-\frac{1}{2} \begin{pmatrix} x \\ y \end{pmatrix}' M^{-1} \begin{pmatrix} x \\ y \end{pmatrix} \right). \quad (41)$$

Using (23), it can be verified that

$$M^{-1} = \begin{pmatrix} R^{-1} + \frac{1}{2h\sigma^2} (I + hB')(I + hB) & -\frac{1}{2h\sigma^2} (I + hB)' \\ -\frac{1}{2h\sigma^2} (I + hB) & \frac{1}{2h\sigma^2} I \end{pmatrix}. \quad (42)$$

We have the following expressions for the target density π and the transition density q with respect to Lebesgue measure:

$$\begin{aligned} \pi(x) &= \frac{1}{(2\pi)^{n/2} \det(V)^{1/2}} \exp(-\tfrac{1}{2} \langle x, V^{-1}x \rangle), \quad \text{and} \\ q(x, y) &= \frac{1}{(2\pi)^{n/2} (2h\sigma^2)^{n/2}} \\ &\quad \times \exp(-|y - (I + hB)x|^2 / (4h\sigma^2)). \end{aligned}$$

Multiplication gives

$$\begin{aligned} \pi(x)q(x, y) &= \frac{1}{(2\pi)^n \det(V)^{1/2} (2h\sigma^2)^{n/2}} \\ &\quad \times \exp \left(-\frac{1}{2} \xi' N^{-1} \xi \right), \end{aligned}$$

where

$$\begin{aligned} \xi &= \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{and} \\ N^{-1} &= \begin{pmatrix} V^{-1} + \frac{1}{2h\sigma^2} (I + hB')(I + hB) & -\frac{1}{2h\sigma^2} (I + hB)' \\ -\frac{1}{2h\sigma^2} (I + hB) & \frac{1}{2h\sigma^2} I \end{pmatrix}. \end{aligned} \quad (43)$$

To satisfy (15), we require that $c\gamma(y, x) + \pi(x)q(x, y) \geq 0$ for all $x, y \in \mathbb{R}^n$ and some constant $c > 0$. We compute

$$\begin{aligned} \pi(x)q(x, y) + c\gamma(y, x) &= \pi(x)q(x, y) + c(f(y, x) \\ &\quad - f(x, y)) \geq \pi(x)q(x, y) - cf(x, y) \\ &= (2\pi)^{-n} (\det V)^{-1/2} (2h\sigma^2)^{-n/2} \exp \left(-\frac{1}{2} \xi' N^{-1} \xi \right) \\ &\quad - c(2\pi)^{-n} \det(M)^{-1/2} \exp \left(-\frac{1}{2} \xi' M^{-1} \xi \right). \end{aligned}$$

By Lemmas 5.6 and 5.7, we have

$$\begin{aligned} \det(M)^{-1/2} &= \frac{1}{(2h\sigma^2)^{n/2}} \det(R)^{-1/2} \\ &\leq \left(\frac{1}{(2h)^{n/2} (\sigma^2)^n (\det V)^{1/2}} \right), \end{aligned}$$

so that, for $c \leq \sigma^n$,

$$\begin{aligned} c\gamma(y, x) + \pi(x)q(x, y) &\geq k \left[\exp \left(-\frac{1}{2} \xi' N^{-1} \xi \right) - \exp \left(-\frac{1}{2} \xi' M^{-1} \xi \right) \right] \\ &= k \exp \left(-\frac{1}{2} \xi' M^{-1} \xi \right) \left[\exp \left(\frac{1}{2} \xi' M^{-1} \xi - \frac{1}{2} \xi' N^{-1} \xi \right) - 1 \right], \end{aligned}$$

where $k = ((2\pi)^{2n} (2h\sigma^2)^n \det V)^{-1/2}$. The last factor is nonnegative for all x, y if and only if $M^{-1} - N^{-1}$ is positive semidefinite. By (42) and (43), we have

$$M^{-1} - N^{-1} = \begin{pmatrix} R^{-1} - V^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Using Lemma 5.8 (iv), we find that for the specified values of h and σ , $R \leq V$ and therefore [by (Horn and Johnson 1990, Corollary 7.7.4)], $R^{-1} - V^{-1} \geq 0$. \square

References

- Chen, T.-L., Hwang, C.-R.: Accelerating reversible Markov chains. Stat. Prob. Lett. **83**(9), 1956–1962 (2013)
- Chen, F., Lovász, L., Pak, I.: Lifting Markov chains to speed up mixing. In: Proceedings of the thirty-first annual ACM symposium on Theory of computing, pp. 275–281. ACM, (1999)
- den Hollander, F.: Large Deviations, Volume 14 of Fields Institute Monographs. American Mathematical Society, Providence, RI (2000)
- Diaconis, P., Holmes, S., Neal, R.M.: Analysis of a nonreversible Markov chain sampler. Ann. Appl. Prob. **10**, 726–752 (2000)
- Diaconis, P.: The Markov chain Monte Carlo revolution. Bull. Am. Math. Soc. **46**(2), 179–205 (2008)
- Diaconis, P., Saloff-Coste, L.: What do we know about the Metropolis algorithm? J. Comput. Syst. Sci. **57**, 20–36 (1998)
- Geyer, C.J., Mira, A.: On non-reversible Markov chains. In: Monte Carlo methods (Toronto, ON, 1998), volume 26 of Fields Inst. Commun., pp. 95–110. Amer. Math. Soc., Providence, RI (2000)
- Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**(1), 97–109 (1970)

- Hwang, C.R., Hwang-Ma, S.Y., Sheu, S.J.: Accelerating Gaussian diffusions. *Ann. Appl. Prob.* **3**(3), 897–913 (1993)
- Horn, R.A., Johnson, C.R.: *Matrix Analysis*. Cambridge University Press, Cambridge (1990)
- Horowitz, A.M.: A generalized guided Monte Carlo algorithm. *Phys. Lett. B* **268**(2), 247–252 (1991)
- Lelièvre, T., Nier, F., Pavliotis, G.A.: Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *J. Stat. Phys.* **152**(2), 237–274 (2013)
- Levin, D.A., Peres, Y., Wilmer, E.L.: *Markov Chains and Mixing Times*. American Mathematical Society, Providence (2009)
- Lancaster, P., Tismenetsky, M.: *The Theory of Matrices*. Computer Science and Applied Mathematics, 2nd edn. Academic Press Inc, Orlando, FL (1985)
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**(6), 1087 (1953)
- Neal, R.M.: Improving Asymptotic Variance of MCMC Estimators: Non-reversible Chains are Better. Technical report, No. 0406, Department of Statistics, University of Toronto, July (2004)
- Neal, R.: MCMC using Hamiltonian dynamics. In: Brooks, S., Gelman, A., Jones, G.L., Meng, X.L. (eds.) *Handbook of Markov Chain Monte Carlo*, pp. 113–162. Chapman & Hall/CRC, Boca Raton (2011)
- Rey-Bellet, L., Spiliopoulos, K.: Irreversible Langevin samplers and variance reduction: a large deviation approach. p. 21, March (2014)
- Roberts, G.O., Rosenthal, J.S.: Optimal scaling of discrete approximations to Langevin diffusions. *J. Royal Stat. Soc.* **60**(1), 255–268 (1998)
- Roberts, G.O., Rosenthal, J.S.: General state space Markov chains and MCMC algorithms. *Prob. Surv.* **1**, 20–71 (2004)
- Roberts, G.O., Tweedie, R.L.: Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**(4), 341–363 (1996)
- Rudin, W.: *Real and Complex Analysis*, 3rd edn. McGraw-Hill Book Co, New York (1987)
- Sun, Y., Gomez, F., Schmidhuber, J.: Improving the asymptotic performance of Markov Chain Monte-Carlo by inserting vortices. In: Lafferty, J., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.), *Advances in Neural Information Processing Systems* 23, pp. 2235–2243. (2010)
- Suwa, H., Todo, S.: Markov Chain Monte Carlo method without detailed balance. *Phys. Rev. Lett.* **120**603, 1–4 (2010)
- Turitsyn, K.S., Chertkov, M., Vucelja, M.: Irreversible Monte Carlo algorithms for efficient sampling. *Physica D* **240**(4–5), 410–414 (2011)
- Vucelja, M.: Lifting—a nonreversible Markov Chain Monte Carlo algorithm. December (2014)
- Wilmer, E.L.: Exact rates of convergence for some simple non-reversible Markov chains. PhD thesis, Harvard University (1999)
- Yosida, K.: *Functional Analysis*. 6th edn (1980)